Georeferencing Historic Collection Data

Caitlin Lara Thorn ‡

‡ Museum für Naturkunde, Berlin, Germany

Corresponding author: Caitlin Lara Thorn (caitlin.thorn@mfn.berlin)

Abstract

Collection data from historic collections often contain vague or non-specific location information of where the specimen was found. Now, during the mass-digitization era of natural history collections, this presents a challenge as we intend to georeference these locations without specific details of where they were found. In a case study at the Museum für Naturkunde Berlin (MfN), a system was developed to georeference these vague locations.

There are three types of geospatial vector data that should be considered in georeferencing: points, lines, and polygons. In most collections, objects were taken from a place that can be represented by a point coordinate (x, y or latitude and longitude). However, if these coordinates were not captured at the time, or information has been lost, making a polygon (i.e. a bounding area) is more appropriate for georeferencing a collection site (Hill 2009). Many databases and standards, however, expect point coordinate information with a field to account for uncertainty. TDWG's Darwin Core Standard includes terms for *decimalLatitude*, *decimalLongitude*, *geodeticDatum*, and *coordinateUncertaintyInMeters*, among additional georeferencing fields (Wieczorek et al. 2012). Therefore, the following process results in relevant information to fulfill this standard.

MfN's Neuroptera collection required the georeferencing of their historic collection of specimens found in Germany. Their data contained verbatim place names for the objects collected between 1758 and 1906. Some of these places were vague and non-specific, referring to entire states or cities, while others were more detailed descriptions. This georeferencing process ultimately resulted in a searchable table of place names in Germany, at different administrative levels. For each, there was a latitude, longitude and uncertainty radius assigned.

Open geospatial data sources of polygon boundaries of Germany's regions at different administrative levels were used as a basis for the project. These were transformed with Q GIS. First, the polygon layer was used to create a circular polygon, encompassing the entire area. Next, a measurement of each circle's circumference was taken and stored in the data's attribute table. Then the centroids of each circle were calculated. These became the latitudes and longitudes for each area. This visualisation is shown in *Fig. 1*. Finally, the

data was tidied, and the radius of the circle was calculated — this became the uncertainty measurement for each area, as it was measured from the centroid of the polygon to the maximum possible distance edge of the polygon.

In total there were six output tables (see *Fig. 2*) — the four administrative levels, and two additional levels for Berlin, which is organised differently (Thorn 2022). These tables allowed a user to search for the verbatim place name they had in the data, and assign coordinates and an uncertainty radius to it.

This method for georeferencing historic locations could be replicated in different countries, eventually creating a comprehensive database that would aid in georeferencing historic (and recent) collections. If this project were to be used more widely, additional outputs could be created using historic boundaries. The outputs from this process can be repeatedly reused, therefore saving collection staff from manually finding coordinates for everything in their collections. While currently the output tables still have to be searched to find relevant data, this may be automated in the future, creating an efficient georeferencing process.

Keywords

location, GIS, coordinates, geography, Germany

Presenting author

Caitlin Lara Thorn

Presented at

TDWG 2022

Conflicts of interest

References

- Hill L (2009) Georeferencing: The Geographic Associations of Information. MIT Press [ISBN 9780262083546]
- Thorn C (2022) Georeferencing Germany: Presenting coordinates and uncertainty measures for administrative boundaries [Dataset]. Museum für Naturkunde Berlin (MfN) - Leibniz Institute for Evolution and Biodiversity Science. URL: https://doi.org/10.7479/672a-zd40

- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715

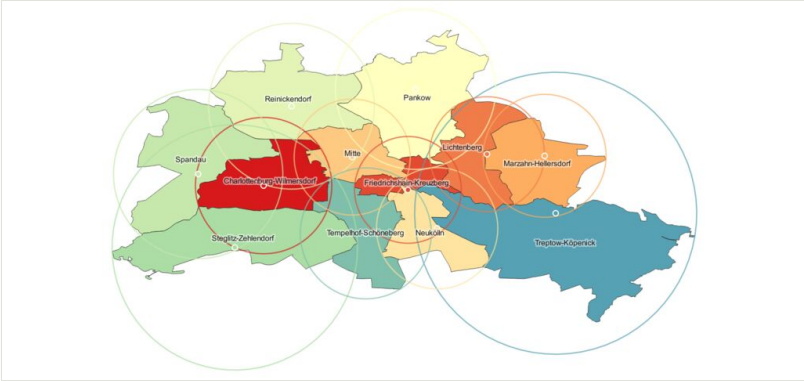

Figure 1.

Process example of Berlin's Bezirk areas (Thorn 2022, license CC BY ND).

Name:	Area/Data Points:	Search in Column:	Notes:
Germany Admin 1	Bundesland / States (16)	NAME_1 (B)	
Germany Admin 2	404 areas	NAME_2 (B)	
Germany Admin 3	4681 areas	NAME_3 (B)	
Germany Admin 4	11303 areas	NAME_4 (B)	*These are the smallest/most specific areas.*
Berlin Bezirk	12 Berlin areas	Bezirk_name (B)	*Admin layers 1-4 do not split Berlin further than state level.*
Berlin Ortsteil	Bezirk broken into 96 more specific areas	ortsteilname (B)	

Figure 2.

Overview and explanation of final output tables (Thorn 2022, license CC BY ND).